Accuracy and Reproducibility of a Multiplex Immunoassay Platform: A Validation Study

Julie A. Bastarache, Tatsuki Koyama, Nancy L. Wickersham, Daphne B. Mitchell, Ray L. Mernaugh, and Lorraine B. Ware

1Division of Allergy, Pulmonary and Critical Care Medicine, Department of Medicine, Vanderbilt University School of Medicine
2Department of Biostatistics, Vanderbilt University School of Medicine
3Department of Biochemistry, Vanderbilt University School of Medicine

Abstract

BACKGROUND—Multiplex immunoassays offer many advantages over singleplex assays for the analysis of multiple analytes in a single sample. We sought to validate a specific multiplex cytokine immunoassay (Human 9-plex cytokine array on the Searchlight® platform by Thermoscientific) prior to use in a large clinical study.

METHODS—We compared spike and recovery of recombinant proteins on the Searchlight® platform to singleplex immunoassays purchased from R&D Systems, measured identical patient samples on the two different platforms, and measured identical patient samples on different days to measure intra- and inter-assay variability.

RESULTS—Assays using the Searchlight® platform had inefficient recovery of spiked recombinant proteins compared to R&D Systems singleplex assays. Assaying identical patients samples on different days on the Searchlight platform had acceptable intra-assay variability (intra-assay coefficient of variation (CV%) range for all analytes of 9.1–13.7) but unacceptably high inter-assay variability (CV% range for all analytes 16.7–119.3) suggesting plate-to-plate variability. Similar assays for individual cytokines on the R&D platform had an intra-assay CV% range of 1.6–6.4 and an inter-assay CV% range of 3.8–7.1. Some deficiencies in Searchlight® assay performance may be due to irregularities in spotting of capture antibodies during manufacturing.

CONCLUSIONS—We conclude that the Searchlight® multiplex immunoassay platform would require extensive additional assay optimization prior to widespread clinical research use.

Keywords

Enzyme linked immunosorbent assay; Biomarkers; Acute lung injury; Acute respiratory distress syndrome

Corresponding author: Julie A. Bastarache, MD, Division of Allergy, Pulmonary and Critical Care Medicine, Vanderbilt University School of Medicine, 1161 21st Avenue South, Suite T-1217 MCN, Nashville, TN 37232. Julie.bastarache@vanderbilt.edu.

Publisher’s Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
INTRODUCTION

Circulating biomarkers are important in the early diagnosis and prognosis of many diseases. Singleplex immunoassays are the most commonly used method for quantifying protein biomarkers in clinical patient samples. However, complex syndromes may require measurement of a panel of biomarkers for diagnosis or prognosis. In addition, identification of novel biomarkers may require the measurement of multiple markers in a single biological sample. Challenges in measuring panels of multiple biomarkers using singleplex immunoassays include low throughput, the large volume of sample required, high reagent and labor costs and handling issues such as the effect of multiple freeze thaw cycles on analyte recovery. In response to the limitations of singleplex immunoassays, multiplex immunoassay technologies to measure biomarkers have become increasingly available. (Tam et al., 2002; Knight et al., 2004; Bozza et al., 2007) In developing multi-plex immunoassays, several factors need to be taken into account. First, antibody characteristics such as antigen specificity (MacBeath, 2002) and binding affinity (Seurynck-Servoss et al., 2007) need to be optimized for each platform. Second, the manufacturing process and antibody immobilization techniques can add variability to the assay. (Ellington et al., 2009) Finally, issues of cross-reactivity and non-specific binding can limit the possible protein combinations that can be multiplexed in a single assay. (Kingsmore, 2006)

Several immunoassay platforms for simultaneous measurement of multiple protein biomarkers are available (reviewed in Ellington et al.). Some platforms use a mixture of antibody labeled fluorescent beads that capture different analytes in a single sample and are assayed using flow cytometry (reviewed in Nolan and Mandy, 2006)). Other platforms use a nitrocellulose membrane or other surface spotted with different capture antibodies that bind analytes in a single sample. The surface is washed and incubated with a detection reagent. (Tam et al., 2002) Finally, 96 well flat bottom (enzyme-linked immunosorbent assay) ELISA plates can be coated with small spots of capture antibodies offering increased sensitivity. (Ekins, 1989) The Searchlight® platform (Thermoscientific, Rockford, IL) uses a black 96-well flat bottom plate coated with spots of different capture antibodies, a cocktail of biotinylated detection antibodies, enzyme-conjugated streptavidin (neutravidin) and a chemiluminescent detection reagent. A specialized (charge-coupled device) CCD imager and analyzer detect the chemiluminescence of each spot in each well and data analysis software uses a standard curve to calculate a concentration of each analyte. The potential advantages of this system over other multiplex immunoassays platforms include the small volume of sample required and increased sensitivity for analyte detection using chemiluminescence. (Fan et al., 2009) However, there have been reports of quality control concerns with this platform. (Ellington et al., 2009) Furthermore, while others have compared the Searchlight platform to other multiplex platforms (Lash et al., 2006) few studies (Backen et al., 2009; Ellington et al., 2009) have done a rigorous intra- and/or inter-assay comparison to test reliability and reproducibility of the Searchlight platform and results from these studies has been inconsistent. None of these studies have combined spike and recovery comparison to single-plex ELISAs, intra- and inter-assay validation of control plasma and patient samples as well as plate image analysis to determine accuracy and reproducibility of the Searchlight multiplex cytokine assay. In response to these concerns, we have carried out a systematic evaluation and validation of the performance of the Searchlight® platform for multiplex immunoassays of human cytokines using both spike and recovery of recombinant proteins and patient samples. We have identified major problems with accuracy and reproducibility with this multiplex platform.
MATERIALS AND METHODS

Assay Kits

These studies were designed to validate a multiplex cytokine immunoassay platform, Searchlight® (Thermo Scientific) prior to use in a large clinical study. For this validation, 9-plex human cytokine assay kits were purchased from Searchlight® for the following analytes: Interleukin (IL)-1α, IL-1β, IL-2, IL-6, IL-8, IL-10, IL-12, interferon (IFN)-γ and tumor necrosis factor (TNF)-α. For comparison, commercially available singleplex cytokine assays for the same analytes were purchased from R & D Systems (Minneapolis, MN). Detection ranges for each analyte on the two different platforms is shown in Table 1.

Patient Samples

Patient samples from two different clinical studies were used. EDTA plasma samples from patients with acute respiratory distress syndrome (ARDS) were used for the plate washing experiments. The patients represent a heterogeneous group of critically ill patients admitted to the intensive care unit. These patients have elevated levels of many of the biomarkers measured, consistent with other critically ill populations reported in the literature. Thus, they represent a good test population and findings from this study are likely to be generalizable to other critically ill patients. For cross platform comparisons, EDTA plasma samples from a large multicenter clinical trial were used. All samples were frozen at −70° at the time of collection in small aliquots to control for freeze/thaw cycles. Each analysis was performed on thawed fresh frozen samples. The institutional review board of Vanderbilt University approved this study.

Plate washing strategies

To determine if plate-spotting irregularities could be corrected with a plate-rinsing step prior to addition of sample, we compared results from unwashed (manufacturers protocol) and prewashed plates. Four lot-matched assay kits (product #84619B, Lot #308584, exp 7-31-2013) and standards were used with each of 2 kits employing a different pre-rinsing strategy on 2 separate days. Plates were either pre-rinsed 3 times with a PBS + 0.1% Tween-20 solution or not pre-rinsed prior to the addition of standards. We then assayed with lot matched assay kits (product #84619B, Lot #308584, exp 7-31-2013) and standards provided by Thermoscientific.

Spike and recovery of recombinant standard proteins

Pooled normal human plasma (U.S Biologicals, Swampscott, MA) was spiked with recombinant standard proteins for recovery analysis on 1) the Searchlight® multiplex platform and 2) R & D Systems (Minneapolis, MN) singleplex ELISA platform. Concentrations were chosen within the standard curve range of both platforms. All samples were frozen at −70° in small aliquots with enough volume for one analysis each to avoid multiple freeze/thaw cycles. Samples were assayed in triplicate, and the concentration of normal pooled plasma run simultaneously for all analytes was used as a zero control and subtracted from the recovery values. Recombinant proteins were obtained from each company and assayed on their respective platforms. Normal plasma was used for the spike and recovery analysis and was diluted 1:5 prior to the addition of the Searchlight® recombinant protein according to manufacturers instructions. In addition, control plasma spiked with recombinant protein from R&D Systems was analyzed on the R&D platform. Standard curves for each platform method were performed according to the respective manufacturers’ instructions using diluents provided in the assay kits.
Platform comparison of human plasma samples

Identical patient samples were assayed on the two different platforms (Searchlight® and R & D Systems) for comparison. Both assays were performed on samples subject to the same number of freeze-thaw cycles and assayed according to each manufacturer’s instructions.

Statistical Analysis

Coefficient of variation (as a percentage) CV% was determined by dividing the standard deviation of the mean divided by the mean value multiplied by 100. Inter-assay CV% was determined for identical sample triplicates analyzed using two different plates from the same lot number on the same platform. Intra-assay CV% was determined for identical sample triplicates.

RESULTS

Plate spotting irregularities

Our first goal was to determine whether there was intra-well uniformity of antibody deposition and luminescence signal in the Searchlight® platform. Our initial studies using a 9-plex human cytokine kit identified multiple problems including halos, comets, and capture antibody drop out. Figure 1A shows the finding of bright rings around the edge of the wells, affecting half of the plate, which we have termed “halos”. These rings were sufficiently luminescent to affect the overall luminescence detection of the individual spots. Figure 1 shows commonly observed irregularities in the multiplex plates with an example of optimal wells (panel 1B) for comparison. A common finding was oversaturation of one spot with bleed-over into an adjacent spot (panel 1C) suggesting that the assay had not been optimized to detect all of the multiplexed analytes at the same sample dilution. We also frequently observed “comets” (panel 1D) trailing from one antibody spot into another. Capture antibody dropout was also observed with loss of capture antibody in some but not other wells from the same plate (panel 1E).

Effect of a pre-rinse strategy on spot irregularities

Because of the halos and comets noted in the initial experiments, we tested whether a pre-rinse of the wells would reduce non-specific background chemiluminescence. Immediately prior to addition of sample, the plates were either pre-rinsed or not pre-rinsed. Pre-rinses were done with PBS/TWEEN 20. The pre-rinsing step did not have any effect on the appearance of comets or halos (data not shown). As there were no differences in washing strategies, we did not do a pre-rinse step for the remainder of the assays and instead followed manufacturer’s instructions.

Inter-assay variability

The plate-to-plate variability was high for the majority of the analytes on the Searchlight® platform. Figure 2 shows the values of all 9 analytes in identical control plasma subjected to the same number of freeze-thaw cycles without the addition of recombinant protein (no calibrator) on seven consecutive, lot-matched plates. There was wide variability in 6 (IL-2, IL-6, IL-8, IL-12, IFNγ and TNF-α) of the 9 analytes. The corresponding CVs for this experiment are also shown.

Recombinant protein spike and recovery

To assess the accuracy of the Searchlight® multiplex assay, pooled normal human plasma diluted 1:5 was spiked with known concentrations of recombinant human proteins obtained from Thermoscientific for analysis on Searchlight® and from R&D Systems for analysis on the R&D platform. Protein concentrations were chosen that were included in the standard
curve range for each assay as antigen excess has the potential to reduce assay signal. The concentration of each analyte in the normal pooled plasma (without spiking) was measured on each plate and used as the zero control by subtracting from all recovery values. Figure 3 shows representative data for spike and recovery of recombinant proteins obtained from both platforms. Recovery of R&D recombinant proteins on the R&D platform was excellent with IFN-γ being the only analyte with sub-optimal recovery. When recombinant proteins obtained from Searchlight® were used for spike and recovery on the Searchlight® platform some of the analytes had acceptable spike and recovery values (IL-10, TNF-α) while others had recovery concentrations that were markedly different from the spiked values (IL-8, IL-6).

For some of the analytes, analysis on the Searchlight® platform yielded negative values for the spiked samples as the recovery values were overwhelmed by the high concentration values of the zero control which was subtracted out of all of the wells. The recovery concentrations for spiked samples analyzed on Searchlight® in two positions on the same plate were similar indicating that there was good intra-assay reproducibility (mean CV% = 4.13)

**Identical human plasma samples assayed on two different platforms**

Identical patient plasma samples from 10 patients enrolled in a clinical trial were assayed on both platforms (Figure 4). Some of the analytes (IL-6 and IL-10) were strongly correlated. Other analytes (TNF-α and IL-1β) had very poor correlation. One potential explanation for discrepancies between measured values for the same samples on the two different platforms may be that for some of the analytes (TNF-α and IL-1β) the analyte levels were below the limit of detection for the R&D assay kit but not for the more sensitive Searchlight assay. Another potential explanation is that the affinity or binding strengths of the antibodies used in the two different assays could be markedly different.

**Identical human plasma samples assayed on two different days**

To determine the intra- and inter-assay reliability of the Searchlight® platform, eight different human plasma samples were run in triplicate on 2 different days. All samples had an identical number of freeze thaw cycles. Table 2 illustrates the findings for all analytes on the two different platforms. In this experiment, intra-assay CVs were calculated from the triplicates and mean CV% and CV% range for each analyte in all 8 patients. In addition, inter-assay CVs were calculated using mean value of each analyte from the two different plates.

**DISCUSSION**

There are several advantages to multiplexing protein immunoassays including cost savings, small volume of sample, and higher throughput. However, in order to be effective tools for research or clinical use, multiplex immunoassay platforms must maintain the same precision and accuracy as singleplex immunoassay measurements. The Searchlight® platform, while theoretically sound, has many practical problems with both accuracy and precision. We have identified many quality control issues with the Searchlight® platform using a rigorous validation procedure. Although there were some advantages to the Searchlight® platform, namely increased sensitivity of all analytes according to manufacturers protocols (Table 1), serious concerns about reliability and reproducibility outweigh these benefits.

The microtiter-based Searchlight® assay platform utilizes antibodies spotted onto microtiter wells to capture antigens from a sample solution. We first identified potential capture antibody spotting irregularities that could result in intra-well spot interference (i.e. halos,
bleed over, and comets). All of these could lead to an artificially high assay signal and an overestimation of the protein analyte concentration. A validation study of two Searchlight vascular multiplex kits did not find problems with signal interference when the concentration of one analyte was varied but did identify some problems with spike and recovery and precision. (Backen et al., 2009) However, this study did not include any images of plates so it is difficult to know if there were problems related to intra-well spot irregularities. In addition to spot irregularities, we noted complete absence of signal for some analytes in replicate wells. Apparently the capture antibody in some wells was not spotted correctly. As such, intra-assay variability was high and there was a potential to underestimate protein concentration based on wells with little or no signal intensity due to capture antibody dropout. Others have found similar issues with plate to plate variability in the Searchlight platform. Ellington, et al analyzed data from 66 Searchlight plates used to measure proteins as part of a large clinical study and found that half of the plates used for their study had quality control issues. (Ellington et al., 2009) Our validation procedure was not designed to test for “under-spotting” for specific analytes but this potential problem could account for some of the wide variability we found in our control plasma replicates (Figure 2) and day-to-day comparisons of the same samples (Table 2). Our attempts to reduce some of the spot irregularities using a pre-rinse strategy did not result in significant improvement.

A second problem that we identified in the Searchlight® platform was inter-assay variability. With both control plasma (Figure 2) in the absence of recombinant protein spiking, and individual patient samples (Figure 4) we found significant inter-assay variability in several of the analytes. The control plasma and samples were subject to the same number of freeze thaw cycles eliminating freeze thaw effects as the source of this day-to-day variability. This degree of plate-to-plate variability is quite concerning when measuring multiple samples from large clinical studies, introducing unintended error in study results. Ellington and colleagues also reported a similar high inter-assay CV problem with the Searchlight platform. (Ellington et al., 2009)

Finally, we identified major problems with spike and recovery of manufacturer supplied recombinant proteins. Using the R&D platform, spike and recovery of all 9 analytes was robust. However, recovery of spiked Searchlight® proteins on the Searchlight® platform was not as good (Figure 3). Because of the high and variable background values in control plasma for many of the analytes on the Searchlight® platform, the recovered values were often below zero after subtracting the zero control value; this issue did not arise in the R&D platform. In a study comparing Searchlight® to R&D systems (as well as a second multiplex platform) Toedter and colleagues (Toedter et al., 2008) also showed significant spike and recovery problems with the Searchlight® platform. In that study variability may have arisen in part due to individual patient factors (single vs. pooled plasma samples) with pooled plasma showing less variability. Although we did not do spike and recovery of individual plasma samples, our study did show significant problems with spike and recovery of recombinant proteins in pooled normal plasma, confirming serious concerns about the reliability of analyte analysis in the Searchlight® assay. The source of variability in the spike and recovery is unclear. Although plate spotting irregularities may have accounted for some of the variability, our study did not identify a single systematic problem resulting in the poor spike and recovery values.

Problems with accuracy and reproducibility are not unique to the Searchlight® multiplex platform. Variability in assay performance has been seen in a plate based multiplex assay developed in Switzerland (Urbanowska et al., 2006) and another developed in the United States (Liew et al., 2007) There are many theoretical limitations to multiplex immunoassay protein measurements including capture antibody cross-reactivity, intra-well interference
issues when dramatically different concentrations are seen in 2 different analytes, and variability in capture antibody spot size or density depending on plate manufacturing process. Unfortunately, the Searchlight® multiplex immunoassay platform appears to suffer from all of these limitations.

In summary, we identified serious problems with inter-assay variability for multiple analytes in the Searchlight multiplex immunoassay platform. Several investigators have outlined validation and quality control criteria that should be met prior to broad research and clinical use. (Ellington et al.; Kricka and Master, 2008) Our findings underscore the need to incorporate rigorous validation protocols prior to use of multiplex platforms in either research or clinical settings.

Abbreviations

- **ELISA**: Enzyme-linked immunosorbent assay
- **CCD**: Charge-coupled device
- **IL**: Interleukin
- **IFN**: Interferon
- **TNF**: Tumor necrosis factor
- **ARDS**: Acute respiratory distress syndrome
- **CV%**: Coefficient of variation percentage

Acknowledgments

Supported by NIH HL081332, HL088263, HL090785, HL103836

REFERENCES


Figure 1. Intra-well spot irregularities

Digital images of representative duplicate wells in a Searchlight® human 9-plex cytokine kit. Panel A shows an ideal well with each spot visible and no overlap of spots. Panel B shows intense signal of one analyte spot (arrow) with the luminescence of that spot bleeding over into an adjacent spot. Panel C shows the comet effect (arrow) with the signal from one spot interfering with an adjacent spot. Panel D shows capture antibody drop-out (arrow) in a pair of duplicate wells. Panel E shows halos around the edge of wells affecting a large portion of the plate (circled area).
Figure 2. Variability in analyte values in control human plasma
Identical aliquots of pooled control human plasma with the same number of freeze thaw cycles were run on seven lot matched Searchlight® plates. The individual biomarker and the corresponding CV% are noted on the x-axis with the value of each of the seven individual marker measurements displayed on the y-axis.
Figure 3. Spike and recovery of R & D proteins on the R & D platform and Searchlight proteins on Searchlight platform
Average values of recombinant proteins spiked into control human plasma for all 9 analytes. Each graph shows the expected concentration of the sample (spike, x-axis) and the measured concentration of the sample (recovery, y-axis). R&D data is plotted on the left and Searchlight data is plotted on the left.
Figure 4. Patient samples assayed on the two platforms
Average values for 5 analytes measured in 10 patient samples on both platforms. R&D values are shown on the x-axis, Searchlight values are shown on the y-axis. Diagonal line represents 100% agreement between the two platforms.
Table 1

Range for each analyte on both platforms.

<table>
<thead>
<tr>
<th>Analyte</th>
<th>R&amp;D Systems (pg/ml)</th>
<th>Searchlight (pg/ml)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-1β</td>
<td>3.9–250</td>
<td>0.39–100</td>
</tr>
<tr>
<td>IL-6</td>
<td>3.12–300</td>
<td>0.39–100</td>
</tr>
<tr>
<td>IL-8</td>
<td>31.2–2000</td>
<td>0.78–200</td>
</tr>
<tr>
<td>IL-10</td>
<td>7.8–500</td>
<td>0.78–200</td>
</tr>
<tr>
<td>IL-12p70</td>
<td>7.8–500</td>
<td>1.2–300</td>
</tr>
<tr>
<td>TNF-α</td>
<td>15.6–1000</td>
<td>4.7–1200</td>
</tr>
<tr>
<td>IL-2</td>
<td>not done</td>
<td>0.78–200</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>not done</td>
<td>0.78–200</td>
</tr>
<tr>
<td>IL-1α</td>
<td>not done</td>
<td>0.78–200</td>
</tr>
</tbody>
</table>
Table 2

Intra-and inter-assay CV% for each platform.

<table>
<thead>
<tr>
<th>Analyte</th>
<th>R&amp;D Systems</th>
<th>Searchlight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intra-Assay CV% Mean (range)</td>
<td>Inter-Assay CV% Mean (range)</td>
</tr>
<tr>
<td>IL-1β</td>
<td>4.8</td>
<td>5.6</td>
</tr>
<tr>
<td>IL-6</td>
<td>2.6</td>
<td>4.5</td>
</tr>
<tr>
<td>IL-8</td>
<td>5.8</td>
<td>7.7</td>
</tr>
<tr>
<td>IL-10</td>
<td>36.9</td>
<td>3.7</td>
</tr>
<tr>
<td>IL-12p70</td>
<td>1.3</td>
<td>4.9</td>
</tr>
<tr>
<td>TNF-α</td>
<td>4.7</td>
<td>5.8</td>
</tr>
<tr>
<td>IL-2</td>
<td>3.1</td>
<td>4.6</td>
</tr>
<tr>
<td>IFN-γ</td>
<td>3.4</td>
<td>6.0</td>
</tr>
<tr>
<td>IL-1α</td>
<td>1.6</td>
<td>3.8</td>
</tr>
</tbody>
</table>